An Optimal Search Process of “Eigen Knots”

for Spline Logistic Regression

John Gao and Cheryl Caswell

Research Department, Point Right

Abstract

The spline regression method usually defines a series piecewise linear variables from a nonlinear continuous variable. In the new spline regression method, the Knots for the piecewise linear variables are derived from an optimal selection processing. And also the categorical variable has been regrouped by using the optimal processing.

First, the all data points have been sorted from low to high based on the value of the variable. Then a portion of the data points in the low end are picked up for the initial regression processing as seed group. The second group as a comparison group is that extra data points near by the seed group are added into the seed group so as to form the combination of these two groups. Then, the statistical analysis is to be conducted to test if the point between seed group and comparison group to be "eigen knot" based on the $X^2$ test on the ratio of likelihood functions. The $X^2$ test will decide if the seed group is significant different from comparison group. If it is true, the largest point in the seed group will be the “eigen knot”. Otherwise, the combination group of the seed and comparison group will be the seed group in the search processing in the next step. The previous search processing will be repeated until all data points of the independent variable have been tested.

We also have conducted a serious comparison studies between the new method with other data mining methods, such as Decision tree and neural network by using Hosmer-Lemeshow good fitness test. The result of the comparison shows that the new method has better result in prediction, efficiency, and stability.
Introduction

The classic regression method is to derive the Global Parameter Model. It is good for Small Data set, or the true function is simple. Recently, the spline method has been developed to be used for data mining on big data set, such as MARS and Data Rewards. The MARS has used the LSE to get the piecewise spline regression functions. The data reward has the spline method, in which all observed points have been separated into the intervals, each interval has the same number of data points. The spline Regression Method is regarded as the Local Mixed Model (Parameter and Non-parameter models) and is good for big data set or true function is complex. The paper has proposed a search processing for the “eigen knots”, which capture the major characteristics of the complex true function of a continuous independent variable. Then the spline functions will be defined from these “eigen knots”. And the spline regression function has

1. More accurate approach to the true function than that in data rewards with less splines
2. the spline functions are derived from maximization of likelihood on binary outcome variable, better than MARS on the LSE approach.

The optimal Spline functions

Given one data set with an independent variable X and dependent variable Y with the N observation points

\[ X : x_0, x_1, x_2, ..., x_N \]

\[ Y : y_0, y_1, y_2, ..., y_N : \text{where } y_i = (0, 1) \text{ for } i=1,2,3,\ldots,N \]
The classic logistic regression is to look for the solution of $A(a_0, a_1, \ldots, a_M)$ with $y = f(x, A)$, the linear function of $f(x, A) = a_0 + a_1 * x$ with Maximize the Likelihood function

$$l(A) = \prod [\exp(f(x, A)) / (1 + \exp(f(x, A)))] = \prod \pi(x, A)$$

For a nonlinear function, we can use a piecewise linear regression function to approach the general nonlinear regression function. The piecewise linear regression function can be defined based on the knots selected from value range of independent variable $X$.

It is assumed that the knots have been selected as

$$X_0, X_1, \ldots, X_k$$

Then the intervals of independent variable for the piecewise function can be defined as

$$R_j = [X_{j-1}, X_j] \quad j = 1, 2, \ldots, k$$

And then piecewise function can be defined as

$$F(x) = f_j(x) \quad \text{if } x \in R_j$$

$$= 0 \quad \text{if } x \notin R_j$$

Where the linear function is that $f_j(x, B_j) = b_{0j} + b_{1j} * x \quad j = 1, 2, \ldots, k$ and $x \in R_j$

Then the likelihood function in the piecewise regression function can be assumed to be

$$l(B) \propto \prod \prod [\exp(f_j(x_i, B_j)) / (1 + \exp(f_j(x_i, B_j)))] = \prod \prod \pi_j(x_i, B_j) = \prod l_j(B_j);$$

where $l_j(B_j)$ is only for $x \in R_j$

Therefore, the maximization of the global likelihood function is achieved by the maximization of all local likelihood functions, i.e.
Max{\(l(B_j)\)} \propto \max\{l_j(B_j)\}, j=1,\ldots,k

To get a best piecewise regression function which can capture the major “eigen knots” of the nonlinear function, we have proposed an optimal processing which can search for these eigen knots with Maximization of the all likelihood functions.

The first step of the optimal search processing of “eigen knots” is to separate all observed points of X into many groups based on the value rank from smallest to largest. For example, \(X: x_0, x_1, x_2, \ldots, x_N\) where \(x_0 < x_1 < x_2 < \ldots < x_N\) can be grouped into \(Z\) groups with the same number of observed points

\(G_1: x_0, \ldots, x_{N_1}\)

\(G_2: x_{N_1+1}, \ldots, x_{N_2}\)

\(\ldots\)

\(G_2: x_{(N_2-1)+1}, \ldots, x_{N_2}\)

here \(x_{N_1}, \ldots, x_{N_2-1}\) are defined as the “Primary knots” to be selected for the final point set. The optimal processing is to select the “eigen Knots” among these “Primary knots” based on statistical analysis.

The first step in the selection processing is from \(G_1\) and \(G_2\). \(G_1\) is called as the seed spline. The combination of \(G_1\) and \(G_2\) is defined as \(G_1^*\). The likelihood functions can be written as

\(l(B_1)\) for \(G_1\) and \(l(B_2)\) for \(G_2\)

and \(l(B_1^*)\) for \(G_1^*\).

The linear logistic regression for \(G_1^*\) is to be conducted to decide the \(B_1: b_0^*, b_1^*\) under the maximization of \(l(B_1^*)\). Then the \(b_0^*, b_1^*\) will be used for the calculation of \(l(B_1)\) with \(b_0^* = b_0^1\) and
\( b^*_1 = b_1^1 \). Then after the \( B_2: b_0^2, b_1^2 \) are decided by maximizing the likelihood function of \( l(B_2) \), we can get the statistic

\[
\text{Statistic } G = -2 \ln \left[ \frac{l(B_1) \cdot l(B_2)}{l(B_1^*)} \right] \sim X^2(2)
\]

Under \( H_0: b_0^1 = b_0^2 \) and \( b_1^1 = b_1^2 \)

If the \( H_0 \) is true, then we can combine \( G_1 \) and \( G_2 \) into \( G_1^* \). In the next step, the seed spline is the combined \( G_1 \) and \( G_2 \). The \( x_{N2} \) is the first “eigen knot” we are looking for. If the \( H_0 \) is not true, The \( x_{N1} \) is the first “eigen knot”. the seed spline in the next step is the \( G_2 \).

If the seed spline is noted as \( G^* \), the next spline is \( G^{**} \). Therefore, in the second step, the next spline is \( G_3 \), the seed spline is \( (G_1, G_2) \) for the \( H_0 \) is true and \( G_2 \) for \( H_0 \) is not true. By repeating the previous processing, we can identify the second “eigen knot”, third “eigen knot” and until all “primary knots” have been tested.

The Chart shows that
1) Class logistic regression functions is the linear and cubic function
2) the data rewards will produce 5 splines to approach the cubic regression function
3) the current optimal regression method only use 3 splines to approach the cubic regression function

Predictive Customer Attrition Model

A SAS Macro program has been developed for the optimal spline regression analysis. The program has been used for many projects in several business fields, such as the predictive model of default risk of private student loans in underwriting, retail bank customer attrition and etc. Here, we mainly discuss the customer attrition model.

The predictive model of customer attrition is to predict who is going to leave the bank according to his activities in the bank data base, including daily transaction, direct ACH deposit, access online accounts and etc. The information can be defined continuous independent variables. Then According to previous optimal processing of univariate analysis, the splines IV’s are defined for all continuous IV’s. Then multi-variable analysis has been conducted for splines IV’s by using stepwise selection method with multiple step selection process.
The final result in the optimal spline logistic regression has the c-statistics of 0.83. The prediction ratio of the top decile vs the average is 4.56.

In addition, we have conducted comparison study on different data mining methods, such as the classic logistic regression, neural network and decision tree (CARTS). The results from Hosmer-Lemeshow good fitness test have shown that these predictions are valid. However, the following table shows the comparison of prediction efficiency. For this data file, the optimal spline logistic regression is the best. The next is the neural network method. And then it is the classic logistic regression. The decision tree is the worst.

Reference.

2) Salford Systems, MARS User Guide. 1999